(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－344076
（P2001－344076A）
(43)公開日　平成13年12月14日(2001.12.14)

(51)Int.Cl.⁷		識別記号	FI		テーマコード（参考）
G06F	3/06	305	G06F	3/06	305C 5B018
		540			540 5B065
	12/16	320		12/16	320L 5D044
		340			340P
G11B	20/10		G11B	20/10	H

審査請求　未請求　請求項の数5　ＯＬ　（全 21 頁）　最終頁に続く

(21)出願番号　特願2000－167484(P2000－167484)

(22)出願日　平成12年６月５日(2000.6.5)

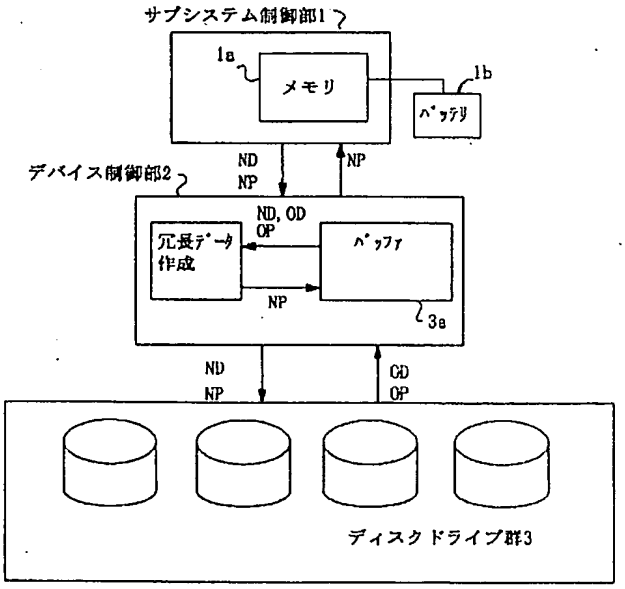
(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　森田　浩文
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　石田　崇
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　100100930
弁理士　長澤　俊一郎　（外１名）

最終頁に続く

(54)【発明の名称】　ディスクアレイ装置

(57)【要約】
【課題】　データの信頼性を維持することができ、性能低下の問題が生ずることが少ないディスクアレイ装置を提供すること。
【解決手段】　ＲＡＩＤ４，５のディスクアレイ装置において、ディスク縮退時、デバイス制御部２で作成した冗長データを、バッテリバックアップされたサブシステム制御部１のメモリ１ａに転送し、書き込みが終了するまではメモリ上に冗長データを保持する。そして、電源瞬断の後の回復時、ディスクドライブ３からデータの読み込みを行わず、メモリ１ａ上の書き込みデータとパリティデータで書き込みを行う。また、ディスク正常時の電源瞬断の後の回復時、書き込み対象となるデータ及びパリティを格納するディスクドライブ以外のディスクドライブから冗長同一グループデータを読出し、それらと書き込みデータから新冗長データを作成し対象ディスクへ書き込む。

本発明の概要を説明する図



1

【特許請求の範囲】

【請求項１】　少なくともメモリがバッテリバックアップされたサブシステム制御部と、

複数のディスクと、該ディスクを制御するデバイス制御部とを備え、

ライトデータを上記複数のディスクのそれぞれに分配して書き込むとともに、該分配された複数のデータから冗長データを生成し、該冗長データを上記複数のディスクのうち、分配したデータを格納したディスク以外のディスクに書き込むディスクアレイ装置であって、

上記ディスクへデータを書き込む際、上記サブシステム制御部のメモリ上に書き込みデータを、書き込み処理が終了まで保持するとともに、上記複数のディスクの内の１乃至複数のディスクが故障したディスク縮退時には、デバイス制御部内のバッファに存在する冗長データをサブシステム側御部のメモリに転送して保持することを特徴とするディスクアレイ装置。

【請求項２】　ディスク縮退時に上記ディスクへデータを書き込む際、書き込み対象となるディスクが故障している場合、ディスクへの書き込みデータと、正常動作しているディスク上の冗長同一グループのデータとから冗長データを生成し、該冗長データを上記サブシステム制御部のメモリに転送することを特徴とする請求項１のディスクアレイ装置。

【請求項３】　電源投入に際し、サブシステム制御部のメモリ中に保持された冗長データを使用して書き込み動作を行うことを特徴とする請求項１または請求項２のディスクアレイ装置。

【請求項４】　電源投入に際し、書き込み対象ディスク以外のディスクに格納された冗長同一グループのデータを用いて冗長データを生成することを特徴とする請求項１または請求項２のディスクアレイ装置。

【請求項５】　電源再投入時に、サブシステム制御部のメモリ内に未書き込みデータが存在するとき、請求項３または請求項４の書き込み動作を行うことを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、複数のディスク装置から構成され、ディスク故障時の修復機能を備えたディスクアレイ装置に関する。

【０００２】

【従来の技術】近年、サブシステム内のコンポーネントを多重化し冗長度をもった装置が実用化されている。この装置は冗長構成と障害の極所化機能により、連続可用性を高めディスク一台の故障時におけるデータの自動修復機能を持つものであり、データの冗長化方法によりＲＡＩＤ０～ＲＡＩＤ５までの６段階に分類される（ＲＡＩＤ：Redundant Array of Inexpensive Disks）。図１４（ａ）にＲＡＩＤ４のシステムの概略を示す。ＲＡＩ

2

Ｄ４は同図に示すように、読み出し／書き込み単位に複数に分配されたデータをそれぞれ格納する複数のデータ用ディスクＤ０，Ｄ２，…と、パリティを生成する手段Ｐと、パリティを格納するディスク装置ＤＰを備え、データの修復情報にはパリティ方式を採用している。

【０００３】データは同図のＡ０，Ａ２，…に示すように複数に分配され（一般には所定長）、分配されたデータＡ０，Ａ２，…はそれぞれデータ用ディスクＤ０，Ｄ２，…に分散して格納され、パリティは一つの専用ディスクＤＰに格納される。なお、以下の説明では、上記のように分配されて異なったディスクに格納されたデータを冗長同一グループデータ、あるいは、単に冗長グループのデータと言い、これらのデータが格納されるディスク群を冗長グループのディスクと言う。また、パリティを冗長データともいう。ディスクに異常が発生した場合、ディスク上のデータは、残りの同一グループデータとパリティ（冗長データ）から再生される。ＲＡＩＤ４は同時に複数の読み出しが可能だが、同時に複数の書き込みはできない。また、更新時には、必ず更新前のデータとパリティを読み出し、更新パリティを作成後書き込むといった余分なアクセスが必要である（これをライトペナルティという）。

【０００４】図１４（ｂ）にＲＡＩＤ５のシステムの概要を示す。ＲＡＩＤ５は、複数に分配されたデータとパリティを格納する複数のディスクＤ１，Ｄ２，…と、パリティを生成する手段Ｐとから構成され、ＲＡＩＤ４と同様、データの修復情報にはパリティ方式を採用している。データは、同図のＡ０，Ａ１，…，Ｂ０，Ｂ１，…に示すように複数に分配され、それぞれディスクＤ１，Ｄ２，…に分散して格納される。また、データＡ０，Ａ１，…のパリティＰＡ、データＢ０，Ｂ１，…のパリティＰＢも各ディスクディスクＤ１，Ｄ２，Ｄ３，…に分散して格納される。ＲＡＩＤ５においても、ＲＡＩＤ４と同様、ディスクに異常が発生した場合、ディスク上のデータは上記同一グループデータとパリティ（冗長データ）から再生される。ＲＡＩＤ５は同時に複数ディスクの読み出し、書き込みが可能であり、更新時には、上記ライトペナルティが発生する。また、パリティ更新中のディスクへは、読み出し／書き込みのアクセスができない。

【０００５】図１５は上記ＲＡＩＤ４又は５を適用可能なディスクアレイ装置における書き込みシーケンス例を示す図であり、同図はＲＡＩＤ４の場合を示している。同図において、１０１はサブシステム制御部、１０１ａはサブシステム制御部内のメモリ、１０２はサブシステム内インタフェース部（以下、インタフェースをＩ／Ｆと略記する）、１０３はデバイス制御部、１０３ａはバッファ、１０４はデバイスＩ／Ｆ部、１０５はディスク群であり、Ｄ０～Ｄ２はデータディスク、Ｐは冗長データを格納する冗長ディスクである。また、同図中のＯＤ

（ＯｌｄＤａｔａ）は更新対象データ（以下では旧データともいう）、ＯＰ（ＯｌｄＰａｒｉｔｙ）は更新対象冗長データ（以下では、旧パリティともいう）、ＮＤ（Ｎｅｗ　Ｄａｔａ）は書き込みデータ、ＮＰ（Ｎｅｗ　Ｐａｒｉｔｙ）は書き込み冗長データ（以下では新パリティともいう）、ＩＰは中間冗長データである。

【０００６】図１５に示すようなディスクアレイ装置おいて、その書き込み動作は、以下のように行われる（図中の(a)～(g)は、以下の(a)～(g)に対応している）。

(a) 書き込みデータＮＤ１をサブシステム制御部１０１のメモリ１０１ａから、デバイス制御部１０３のバッファ１０３ａへ転送する。

(b) 書き込み対象となるディスク内のデータＯＤ１をバッファ１０３ａへ読み出す。

(c) 書き込み対象データの冗長グループの冗長データＯＰをバッファ１０３ａへ読み出す。

(d) ＯＤ１とＯＰを排他的論理和演算することにより、中間冗長データＩＰを生成する。

(e) ＮＤ１とＩＰを排他的論理和演算することにより新冗長データＮＰを生成する。

(f) ＮＤ１をディスク１０５へ書き込む。

(g) ＮＰをディスク１０５へ書き込む。

この際(a)～(c)及び(e)～(f)範囲内での順序は固定されなくとも書き込み動作は実現可能とされる。

【０００７】

【発明が解決しようとする課題】上記のような書き込み動作を行うシステムにおいて停電等による瞬断が発生した場合のデータの信頼性の維持には以下のような方法が考えられている。

（１）装置全体のバッテリバックアップシステムによるサブシステムの動作の継続。

（２）不揮発性メモリによる書き込みデータの維持

上記（１）においては装置に供給される電源が断たれた場合においてもサブシステムの動作は継続されるためそのデータは保証される。上記（２）においては書き込みデータがメモリ上に残存するため電源再投入後に再度ディスクへの書き込み動作を行うことで殆どの場合リカバリ可能である。しかしながら、上記（１）においてはサブシステム全体をバックアップするためには大容量なバッテリを必要とし、実装上その占める割合は非常に大きなものとなる。また上記（２）においては電源断前に書き込みが行われている途中でかつ冗長データが書き込み途中であった場合で、そのＲＡＩＤ内が電源断前に縮退モード（少なくとも一台のディスクが故障時）にある、又は電源再投入後にＲＡＩＤ内の縮退モードに移行した場合には、その冗長グループの冗長性は失われ、故障ディスクのデータの復元又はその冗長グループのデータの書き込みは不可能（正しく行われない）となる。このような状態はＷｒｉｔｅ　Ｈｏｌｅとも表現されている。

【０００８】このような問題に対応するため、メモリ１０１ａ上に冗長データを常時管理し、また書き込みの進行状況をメモリ１０１ａ上に記録しその進行状況によりその冗長データを使用してリカバリを行うようにし、上記状態を避けることも考えられてきているが、書き込み時に常に冗長データの転送が必要となるなどの性能低下が問題となる。本発明は上記事情に鑑みなされたものであって、本発明の目的は、縮退モードであったり、電源断の後の電源再投入時であっても、データの信頼性を維持することができ、性能低下の問題が生ずることが少ないディスクアレイ装置を提供することである。

【０００９】

【課題を解決するための手段】図１は本発明の概要を説明する図である。同図に示すように、本発明は、少なくともメモリ１ａがバッテリ１ｂによりバックアップされたサブシステム制御部１と、複数のディスクからなるディスクドライブ群３と、ディスクを制御するデバイス制御部２とを備え、ライトデータを上記複数のディスクのそれぞれに分配して書き込むとともに、該分配された複数のデータから冗長データを生成し、該冗長データを上記複数のディスクのうち、分配したデータを格納したディスク以外のディスクに書き込む前記ＲＡＩＤ４．５のディスクアレイ装置において、以下のモードを設け、前記正常時、ディスク縮退時、電源瞬断後の回復時の書き込み時において、各状態に応じたモードで動作させる。

１）正常時には、更新対象データＯＤと更新対象冗長データＯＰと書き込みデータＮＤとから新冗長データＮＰを生成し、書き込みデータＮＤと新冗長データＮＰを対象ディスクへ書き込む。

２）ディスク縮退時には、データの書き込みに際し、デバイス制御部で作成した冗長データＮＰを、バッテリバックアップされたサブシステム制御部のメモリに転送する。そして、書き込みが終了するまではメモリ上に冗長データＮＰを保持する。

３）電源瞬断の後の再書き込み時、その対象となるデータ及びパリティを格納するディスクドライブ以外のディスクドライブから冗長同一グループ内のデータＯＤを読出し、それらとメモリ内の書き込みデータＮＤから新冗長データＮＰを作成し、書き込みデータＮＤと新冗長データＮＰを対象ディスクへ書き込む。

４）ディスク縮退時に、電源瞬断の後の再書き込みを行う場合、ディスクドライブからデータの読み込みを行わず、バッテリバックアップされたメモリ上の書き込みデータＮＤとパリティデータＮＰで書き込みを行う。

以上のように、本発明においては、縮退モードのとき、冗長データＮＰを電源バックアップされたメモリ１ａに転送するようにしたので、縮退モード時における電源瞬断後の再書き込み時にも、正しい書き込みを行うことができる。このため、常時、冗長データＮＰをメモリに転送する必要がなく、性能劣化を最低限に抑止でき、か

つ、高い信頼性を維持することができる。また、ＲＡＩＤグループ内の他ディスクの冗長同一グループデータＯＤと、メモリ内の書き込みデータＮＤから新冗長データＮＰを生成することにより、システム電源断後の電源再投入時、ディスク上のデータが正しくない、もしくは信用できない場合であっても、正しい書き込みを行うことができる。

【００１０】

【発明の実施の形態】図２に本発明が適用可能なディスクアレイ装置の概略構成例を示す。図２は大規模なディスクアレイ装置に用いられる構成を示しており、１１はサブシステム制御部であり、ＭＰＵ、メモリ等を含む。１２は上記サブシステム制御部１１と他の制御部とを接続するための内部Ｉ／Ｆであり、複数のプロセッサを有するデバイス制御部１３が接続される。デバイス制御部１３は上記サブシステム制御部１１の指示に従い、ディスクドライブ群１５を制御する制御部であり、バッファ１３ａ、ＭＰＵ１３ｂ、デバイスＩ／Ｆ部１４を制御するためのコントローラ、ＲＡＩＤ４，５動作において必要な排他的論理和（Ｅｘｃｌｕｓｉｖｅ　ＯＲ：ＸＯＲ又はＥＯＲと記述されることもある）演算機能などを有する。１４はデバイス制御部とディスクドライブを接続するためのデバイスＩ／Ｆ部であり、この構成例はシリアルＩ／Ｆのイメージであるが、パラレルＩ／Ｆであっても良い。１５はデータが格納されるディスクドライブ群である。

【００１１】図３は小規模なディスクアレイサブシステムに用いられる構成を示しており、２１はサブシステムを制御するサブシステム制御部であり、サブシステム制御部２１は図２に示したデバイス制御部の機能を兼ね、前記したメモリに加え、バッファ部２１ａとＸＯＲ演算機能を有する。１４，１５はそれぞれ図２に示したデバイスＩ／Ｆ部、ディスクドライブ群である。図４に図２に示す装置のハードウェア構成例を示す。同図において、サブシステム制御部１１はチャネルＩ／Ｆ部を介して上位装置に接続されており、サブシステム制御部１１はメモリ１１ａ、ＭＰＵ１１ｂ、バスインタフェース部１１ｃを備えており、上記ＭＰＵ１１ｂはメモリ１１ａに格納されているプログラムに従って動作する。また、メモリ１１ａには、プログラムの他に、転送データや制御データが格納される。上記サブシステム制御部１１の少なくともメモリ１１ａはバッテリ１１ｄによりバックアップされ、電源断時にも格納されたデータ等は保持される。なお、メモリ１１ａをバッテリバックアップしても長時間の電源断には対応することができないので、例えば、図示しない特定のディスクドライブをバッテリバックアップし、電源断時にメモリ１１ａ内のユーザデータ及び制御情報（冗長データを含む）を上記ディスクに書き込み、電源再投入後にメモリ上に再度展開するように構成してもよい。

【００１２】１３はデバイス制御部であり、デバイス制御部１３は、バッファ１３ａ、ＭＰＵ１３ｂ、上記ＭＰＵ１３ｂを動作させるプログラム等を格納したメモリ１３ｃ、バスインタフェース部１３ｄを備えている。前記したＸＯＲ演算は、例えば上記ＭＰＵ１３ａにより実行される。上記サブシステム制御部１１とデバイス制御部１３はバスＢＵＳを介して接続されており、前記図２に示したサブシステム内Ｉ／Ｆ部１２は、上記バスＢＵＳ、バスインタフェース部１１ｃ、バスインタフェース部１３ｄ等から構成され部分に対応する。デバイス制御部１３は前記したようにデバイスＩ／Ｆ部１４を介してディスクドライブ群１５に接続される。

【００１３】次に、図２に示すディスクアレイ装置を前述のＲＡＩＤ４、５を適用した場合について、本実施例の動作を説明する。なお、以下の説明ではＲＡＩＤ４について説明するが、ＲＡＩＤ５の場合についても同様に適用することができる。また、以下の実施例では、図２に示すシステムについて説明するが、本発明は図３のシステムにも同様に適用することができる。図２、図４において、本実施例のサブシステム制御部１１はディスクドライブ群内で構成されるＲＡＩＤの状態を正常／縮退／データ喪失（同一グループのデータの内、２つのディスクのデータが読めないもしくは信用できない場合）に区別して個々に管理し、その状態を参照してディスクドライブ群への書き込み命令を実行する。

【００１４】また、このサブシステム制御部１１はデバイス制御部１２に書き込みを指示する際、以下のモードを指定することができ、デバイス制御部１２はそれを実行する機能を持つ。

(1) モード１

更新対象データＯＤと、更新対象冗長データＯＰを読出し、排他的論理和をとり中間冗長データＩＰを生成し、さらに、書き込みデータＮＤとの排他的論理和をとり、書き込み冗長データＮＰを生成し、ＮＰとＮＤの対象ディスクへの書き込みを行う。この「モード１」は正常時に選択され動作は前記図１５で説明した通りである。

(2) モード２

デバイス制御部１３がＮＰを作成した際、それをサブシステム制御部１１のメモリ１１ａ上へ転送し、転送後に対象ディスクへの書き込みを行う。このモード２は、後述するようにディスク縮退状態の書き込み時に選択される。

【００１５】(3) モード３

ＯＤ、ＯＰをディスクから読み出すことなく、冗長同一グループの他のディスク内の上記ＯＤと同一のアドレスからデータを読出しそれらとメモリ上のＮＤからＮＰを生成し、対象ディスクへの書き込みを行う。このモード３は、後述するように、ディスクが縮退状態でないとき、書き込み中に予期しない電源断が生じ、電源再投入後にサブシステム制御部１１のメモリ１１ａに未書き込

みデータが存在している場合の再書き込み時に選択される。

(4) モード４

ＯＤ、ＯＰをディスクから読み出すことなく、メモリ上にあるＮＰとＮＤの対象ディスクへの書き込みを行う。このモード４は、後述するようにディスク縮退時、書き込み中に予期しない電源断が生じ、電源再投入後にサブシステム制御部１１のメモリ１１ａに未書き込みデータが存在している場合の再書き込み時に選択される。

【００１６】上記モードの切り替えは次のように行われる。ディスクの一台が故障し、デバイス制御部１３からサブシステム制御部１１が異常終了通知を受けると、サブシステム制御部１３は、特定ディスクへのアクセスが不能で動作継続不可能である場合、縮退状態と認識し、書き込み時のモードを正常モードから縮退モードに移行する。また、サブシステム制御部１１の書き込み命令に対するデバイス制御部１３からの正常応答の有無を判断し正常応答が無いままサブシステム制御部１１が装置電源断を認識した場合、その書き込み処理を仕掛かり中と判断し、電源再投入時に、未書き込みデータがメモリ１１ａ上に存在している場合、上記モード３（ディスク正常時），モード４（ディスク縮退時）で書き込みを行う。

【００１７】図５に、上記正常時のモード（モード１）から縮退時のモード（モード２）への移行のフローチャートを示す。図５において、上位装置からＲＡＩＤへのアクセスがあり、正常終了であれば、正常モードのまま次のＲＡＩＤへのアクセスを待つ。正常終了でない場合には、特定ディスクへのアクセスが不能であるのかを調べる。特定ディスクへのアクセスが不能である場合には、縮退モードに移行する。また、特定ディスクへのアクセスが不能でない場合には、リトライを行い、正常終了すれば正常モードのまま、次のＲＡＩＤへのアクセスを待つ。正常終了しない場合には、所定回数に達するまでリトライを行い、所定回数に達すると、エラー処理を行う。

【００１８】次に、本実施例のディスクアレイシステムの動作について説明する。

（１）正常時

対象となるＲＡＩＤが正常な状態ではサブシステム制御部は、前記「モード１」での書き込み動作をデバイス動作に対し指示を行う。この場合の動作は、図１５で説明したのと同じである。図６に、正常時における通常の書き込み動作のフローチャートを示す。同図において、まず、書き込みデータＮＤ１をメモリ１１ａに保持したまま、バッファ１３ａに転送する（ステップＳ１）。ついで、書き込み対象となっている旧データ（更新対象データ）ＯＤ１をディスク１５からバッファ１３ａに読み込む（ステップＳ２）。次に、書き込み対象の冗長データ（更新対象冗長データ）ＯＰをディスク１５からバッファ１３ａへ読み込む（ステップＳ３）。

【００１９】そして、ＯＤ１とＯＰの排他的論理和（ＸＯＲ）を演算し、中間パリティ（中間冗長データ）ＩＰを作成し（ステップＳ４）、さらに、ＮＤ１とＩＰの排他的論理和（ＸＯＲ）を演算し、新パリティ（書き込み冗長データ）ＮＰを作成する（ステップＳ５）。以上のようにして新パリティＮＰが作成されたら、書き込みデータＮＤ１、新パリティＮＰのディスク１５への書き込みを開始する（ステップＳ６）。そして、書き込みが終了するまで待ち（ステップＳ７）、書き込みが正常終了したらメモリ１１ａ上のＮＤ１を削除して（ステップＳ９）処理を終了する。また、書き込みが正常終了しなかった場合には、ステップＳ８からステップＳ１０に行き、リトライ回数が所定回数になったかを調べ、所定回数になっていなければ、ステップＳ６に戻り、書き込みをリトライする。そして、所定回数リトライしても正常終了しない場合にはエラー処理を行う（ステップＳ１１）。

【００２０】（２）ディスクドライブ群のあるディスクドライブに異常が発生した場合（ディスク縮退時）

ディスクドライブ群のあるディスクドライブに何らかの異常が発生した場合は図７に示すように動作する。サブシステム制御部１１からの読出し／書き込み又はその他のディスクアクセス系のコマンドに対し、ディスクドライブ群１５のあるディスクドライブ（例えば図７におけるディスクＤ１）に何らかの異常が発生した場合、デバイス制御部１３はサブシステム制御部１１へ、異常終了通知を行う（図７の(a)）。通知を受けたサブシステム制御部１１は、その内容を判断してディスクの故障による動作継続不可能である場合、内部の構成変更情報の変更（ＲＡＩＤ状態を縮退状態に変更する）を行い（図７の(b)）、そのディスクドライブを切り離す命令を出し（図７の(c)）、縮退モードに移行する。

【００２１】この縮退モードにあるＲＡＩＤに対する書き込み動作を図８のフローチャートにより説明する。書き込みデータＮＤ２をサブシステム制御部１１のメモリ１１ａに保持したまま、書き込みデータＮＤ２を、デバイス制御部１３のバッファ１３ａに転送する（ステップＳ１）。ついで、ＮＤ２が故障したディスクに書き込むデータであるかを調べ（ステップＳ２）、故障ディスクに書き込むデータでない場合には、書き込み対象となっている旧データ（更新対象データ）ＯＤ２をディスク１５からバッファ１３ａに読み込み（ステップＳ３）、更新対象冗長データ（ＯＰ）をディスク１５からバッファ１３ａに読み込む（ステップＳ４）。ついで、旧データＯＤ２と更新対象冗長データＯＰの排他的論理和（ＸＯＲ）を作成しＩＰとし（ステップＳ５）、書き込みデータＮＤ２とＩＰの排他的論理和（ＸＯＲ）を作成しＮＰとする（ステップＳ６）。

【００２２】また、ＮＤ２が故障したディスクに書き込

むデータである場合には、更新対象データＯＤ２の冗長同一グループデータＯＤ０、ＯＤ１をディスク１５から読み込み（ステップＳ７）、ＯＤ０とＯＤ１と書き込みデータＮＤ２との排他的論理和を求めて新パリティＮＰとする（ステップＳ８）。上記のようにして新パリティＮＰが作成されたら、ＮＰをサブシステム制御部１１のメモリ１１ａに転送する（ステップＳ９）。ついで、正常終了するまで待ち、書き込みデータＮＤ２、新パリティＮＰのディスク１５への書き込みを開始する（ステップＳ１０，Ｓ１１）。そして、書き込みが終了するまで待ち（ステップＳ１２）、書き込みが正常終了したらメモリ１１ａ上のＮＤ２を削除して処理を終了する（ステップＳ１４）。また、書き込みが正常終了しなかった場合には、ステップＳ１３からステップＳ１５に行き、リトライ回数が所定回数になったかを調べ、所定回数になっていなければ、ステップＳ１１に戻り、書き込みをリトライする。そして、所定回数リトライしても正常終了しない場合にはエラー処理を行う（ステップＳ１６）。以上のように、ディスク縮退時には、作成した新冗長データＮＰをサブシステム制御部１１のメモリ１１ａに転送し保持する（前記モード２の動作）。

【００２３】図９に、上記した書き込みデータＮＤ２が故障したディスクに書き込むデータでない場合の動作シーケンス例を示す。サブシステム制御部１１は書き込み動作をデバイス制御部１３に指示する前に該当するＲＡＩＤの状態を確認する。この場合、その書き込み対象となるＲＡＩＤの状態は縮退とされているのでサブシステム制御部１１はデバイス制御部１３に対して次のように書き込みを指示する。

(a) 書き込みデータＮＤ２をメモリ１１ａに保持したまま、メモリ１１ａからバッファ１３ａへ転送する。

(b) 書き込み対象となるディスク内の旧データＯＤ２をバッファ１３ａへ読み出す。

(c) 書き込み対象部分が属する冗長グループの冗長データＯＰをバッファ１３ａへ読み出す。

(d) ＯＤ２とＯＰをＸＯＲ演算することにより、中間冗長データＩＰを生成する。

(e) ＮＤ２とＩＰをＸＯＲ演算することにより新冗長データＮＰを生成する。

(f) ＮＰをサブシステム制御部１１のメモリ１１ａに書き込む。

(g) ＮＤ２，ＮＰをディスク１５へ書き込み、メモリ１１ａ上のＮＤ２を削除する。

【００２４】図１０に上記した書き込みデータＮＤ２が故障したディスクに書き込むデータの場合の動作シーケンス例を示す。

(a) 書き込みデータＮＤ２をメモリ１１ａに保持したまま、メモリ１１ａからバッファ部１３ａへ転送する。

(b) (c) ＯＤ２、ＯＰをディスクから読み出すことなく、冗長グループの他のディスク内の旧データＯＤ０，ＯＤ１をバッファ部１３ａへ読み出す。

(d) ＯＤ０，ＯＤ１，ＮＤ２をＸＯＲ演算することにより、新冗長データＮＰを生成する（中間冗長データＩＰの作成については同図では省略されている。以下同様にＩＰの作成についての説明は省略する）。

(e) ＮＰをサブシステム制御部１１のメモリ１１ａに書き込む。

(f) ＮＤ２，ＮＰを書き込み対象ディスク１５へ書き込み、メモリ１１ａ上のＮＤ２を削除する。

【００２５】（３）電源断後の回復時の動作

次に、書き込み動作中に瞬断等による予期せぬ電源断の場合を考える。前記「モード１（正常時のモード）」で、書き込み動作中に予期せぬ電源断が生じた際、サブシステム制御部１１のメモリ１１ａの電源のみがバックアップされているとした場合、ディスクドライブ群１５への対象ディスクに対する書き込みは中断される。すなわち、サブシステム制御部１１の書き込み命令に対するデバイス制御部からの正常応答の有無を判断し正常応答が無いままサブシステム制御部１１が装置電源断を認識した場合、その書き込み処理を仕掛かり中と判断し、対象ディスクに対する書き込みは中断される。その後、電源が再投入された後にメモリ１１ａ上に未書き込みのデータが存在している場合には書き込みは再試行される。ここで、ディスク縮退状態でない場合（正常時）であっても、再度「モード１（正常時のモード）」で書き込みを行った場合、電源前の書き込みは中断しているため旧データＯＤ（更新対象データ）の読み込み時のデータは正しくない可能性があり使用することはできない。そこで電源投入時に未書き込みのデータが存在している場合、サブシステム制御部１１は前記「モード３」で書き込みを行うようにデバイス制御部１３に指示を出す。

【００２６】また、ディスクが縮退状態にあるとき、その書き込みは前記「モード２」で行われる。「モード２」で書き込み動作中に、予期せぬ電源断が生じた場合、サブシステム制御部１１のみ電源がバックアップされているとした場合、前記したようにディスクドライブ１５のそのディスクに対する書き込みは中断される。その後、電源が再投入された後にメモリ１１ａ上に未書き込みのデータが存在している場合には書き込みは再試行されるが、このとき再度「モード２」で書き込みを行った場合、電源前の書き込みは中断しているため、更新対象データ（旧データ）ＯＤの読み込み時のデータは正しくない可能性があり使用することはできない。さらに、前記「モード３」で書き込みを再試行した場合、そのＲＡＩＤはすでに縮退状態にあるためＰａｒｉｔｙの生成は不可能となる。しかし、瞬断前の書き込みは前記「モード２」で行われているためメモリ内に未書き込みの新冗長データＮＰが保持されている。そこで電源投入時にメモリ１１ａ内に未書き込みのデータおよび未書き込みのＮＰが存在している場合には、サブシステム制御部１

１は前記「モード４」で書き込みを行うようデバイス側御部１３に指示を出す。

【００２７】上記電源瞬断後の回復時の書き込み動作について、図１１のフローチャートにより説明する。電源瞬断後の回復時、メモリ１１ａ上に書き込みデータＮＤ２が登録されているかを調べ（ステップＳ１）、登録されていない場合には処理を終了する。また、登録されている場合には、ＮＤ２の属する冗長グループが縮退モードになっているかを調べる（ステップＳ２）。縮退モードになっていない場合には、書き込み対象となっている更新対象データＯＤ２と冗長同一グループのデータＯＤ０、ＯＤ１をディスク１５からバッファ１３ａに読み込む（ステップＳ３）。また、メモリ１１ａからバッファ１３ａに書き込みデータＮＤ２を転送する（ステップＳ４）。ついで、上記ＯＤ０とＯＤ１と書き込みデータＮＤ２の排他的論理和（ＸＯＲ）を作成し、新冗長データＮＰとする（ステップＳ５）。一方、縮退モードの場合には、ステップＳ６において、メモリ１１ａ上に新冗長データＮＰが保持されているかを調べ、保持されている場合には、書き込みデータＮＤ２、ＮＰをバッファ１３ａに転送する。また、保持されていない場合にはステップＳ３に行く。ついで、書き込みデータＮＤ２、新パリティＮＰのディスク１５への書き込みを開始する（ステップＳ８）。

【００２８】そして、書き込みが終了するまで待ち（ステップＳ９）、書き込みが正常終了したらメモリ１１ａ上のＮＤ２を削除して処理を終了する（ステップＳ１０，Ｓ１１）。また、書き込みが正常終了しなかった場合には、ステップＳ１０からステップＳ１２に行き、リトライ回数が所定回数になったかを調べ、所定回数になっていなければ、ステップＳ８に戻り、書き込みをリトライする。そして、所定回数リトライしても正常終了しない場合にはエラー処理を行う（ステップＳ１３）。

【００２９】図１２に縮退状モードでない場合における電源瞬断後の回復時の書き込み動作シーケンス例を示す。

(a) 書き込みデータＮＤ２をメモリ１１ａからバッファ部１３ａへ転送する。

(b) (c) 冗長グループ内のＯＤ２と同一のアドレスから冗長同一グループのデータＯＤ０，ＯＤ１をバッファ部１３ａへ読み出す。

(d) 上記冗長同一グループデータＯＤ０、ＯＤ１、ＮＤ２をＸＯＲ演算し、新冗長データＮＰを作成する。

(e) ＮＤ２，ＮＰをディスクへ書き込む。

【００３０】図１３に縮退モードにおける電源瞬断後の回復時の書き込み動作シーケンス例を示す。

(a) 書き込みデータＮＤをメモリ１１ａからバッファ部１３ａへ転送する。

(b) ＮＰをメモリ１１ａからバッファ部１３ａへ転送する。

(c) ＮＤをディスクへ書き込む。

(d) ＮＰをディスクへ書き込みむ。

【００３１】

【発明の効果】以上説明したように、本発明をＲＡＩＤ４，５が適用可能なディスクアレイ装置に適用することにより以下の効果を得ることができる。

（１）縮退モードのときに冗長データを電源バックアップされたメモリに転送するようにしたので、従来例のように常時、冗長データをメモリに転送する場合と比べ、性能劣化を最低限に抑止でき、かつ、高い信頼性を維持することができる。

（２）縮退モードにおいて、システム電源断後の電源再投入時の書き込み動作時にメモリ上の冗長データを使用することでディスク上への冗長データの書き込み途中で有る無いに関わらず正しい書き込みを行うことができる。

（３）ＲＡＩＤグループ内の他ディスクの同一グループデータと、メモリ内の書き込みデータから新冗長データを生成することにより、縮退時、あるいは、ディスク上のデータが書き込み途中で有る無いに関わらず正しい冗長データを作成することができる。

特に、ディスク正常時のシステム電源断後の電源再投入時に、ＲＡＩＤグループ内の他ディスクの同一グループデータと、メモリ内の書き込みデータから新冗長データを生成することにより、ディスク上のデータが正しくない、もしくは信用できない場合であっても、正しい書き込みを行うことができる。

【図面の簡単な説明】

【図１】本発明の概要を説明する図である。

【図２】本発明が適用可能なディスクアレイ装置の概略構成例を示す図（１）である。

【図３】本発明が適用可能なディスクアレイ装置の概略構成例を示す図（２）である。

【図４】本発明の実施例のハードウェア構成例を示す図である。

【図５】縮退モードの設定処理のフローチャートである。

【図６】正常時における書き込み動作のフローチャートである。

【図７】ディスクドライブ群のあるディスクドライブに何らかの異常が発生した場合の動作を説明する図である。

【図８】縮退モードのときの書き込み動作を示すフローチャートである。

【図９】縮退モードの書き込み動作シーケンス例（１）を示す図である。

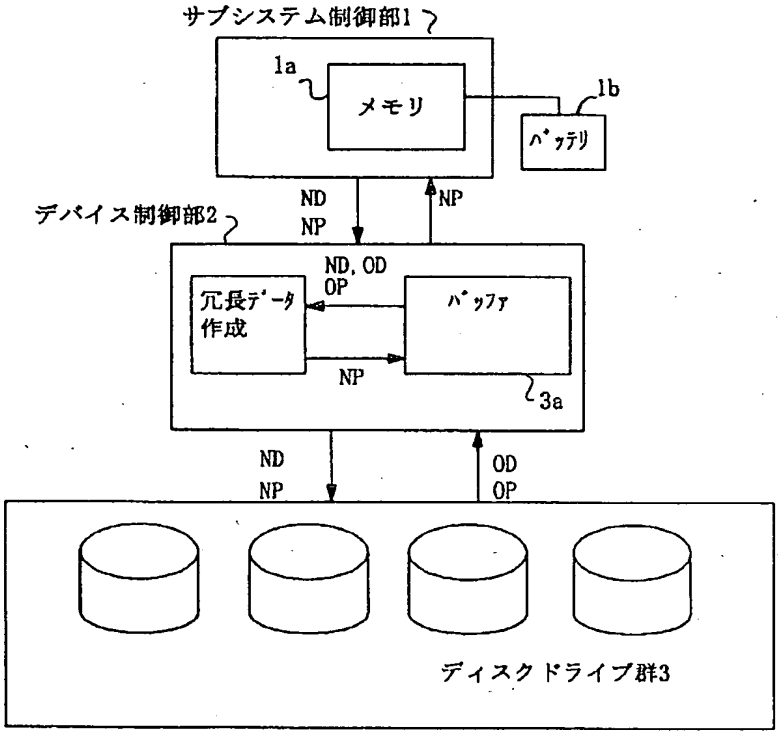
【図１０】縮退モードの書き込み動作シーケンス例（２）を示す図である。

【図１１】電源瞬断後の回復時の書き込み動作を示すフローチャートである。

【図１２】電源瞬断後の回復時の書き込み動作シーケンス例（縮退モードでないとき）を示す図である。
【図１３】電源瞬断後の回復時の書き込み動作シーケンス例（縮退モード時）を示す図である。
【図１４】ＲＡＩＤ４，５を説明する図である。
【図１５】ディスクアレイ装置における書き込み動作シーケンス例を示す図である。
【符号の説明】
１　　サブシステム制御部
１ａ　　メモリ
１ｂ　　バッテリ
２　　デバイス制御部
３　　ディスクドライブ群
１１　　サブシステム制御部
１１ａ　メモリ
１１ｂ　ＭＰＵ
１２　　サブシステム内インタフェース部
１３　　デバイス制御部
１３ａ　バッファ
１３ｂ　ＭＰＵ
１４　　デバイスインタフェース部
１５　　ディスクドライブ群

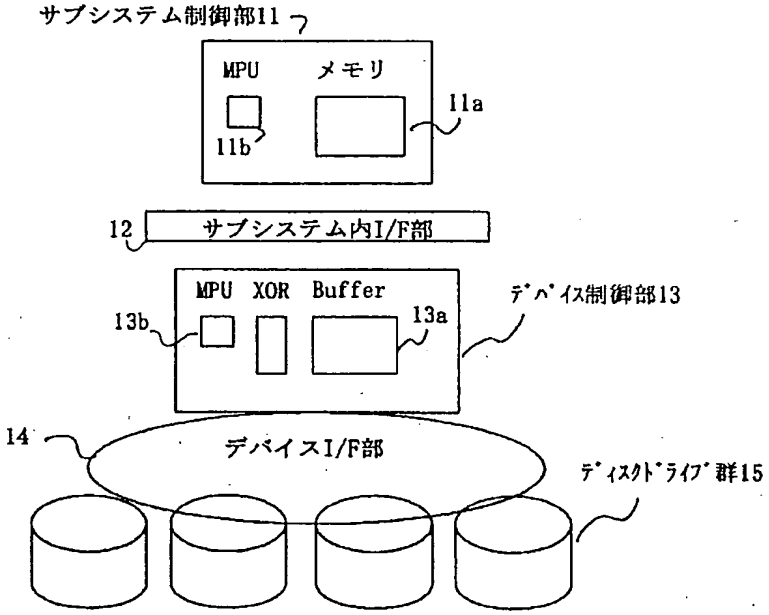
【図１】

本発明の概要を説明する図



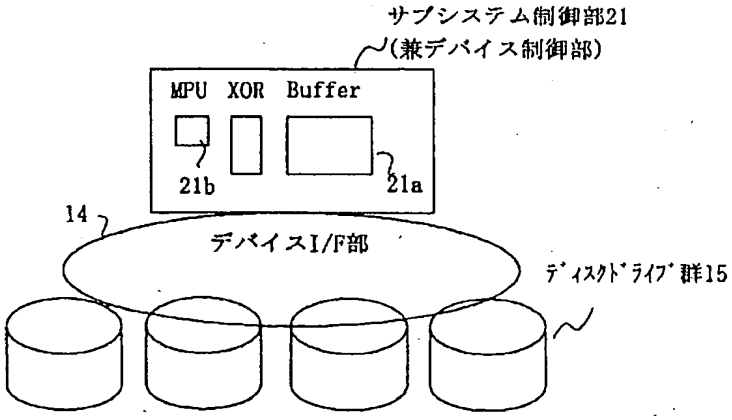
【図２】

本発明が適用可能なディスクアレイ装置の概略構成例（１）を示す図



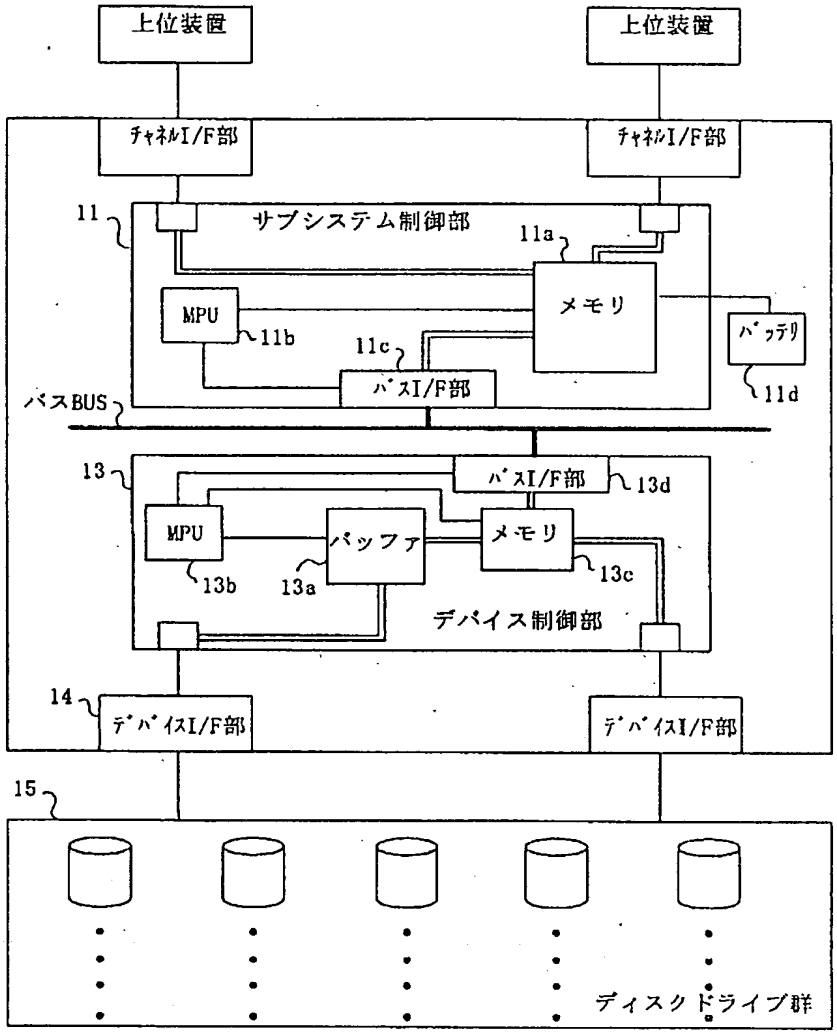
【図３】

本発明が適用可能なディスクアレイ装置の概略構成例（２）を示す図



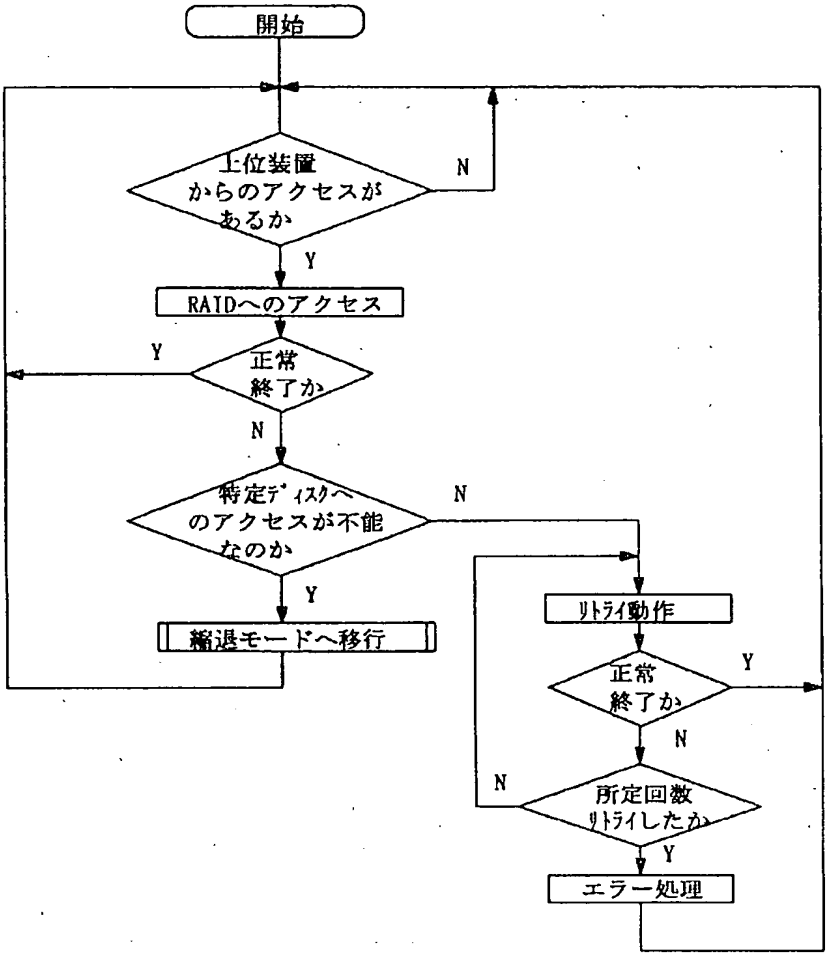
【図４】

本発明の実施例のハードウェア構成例を示す図



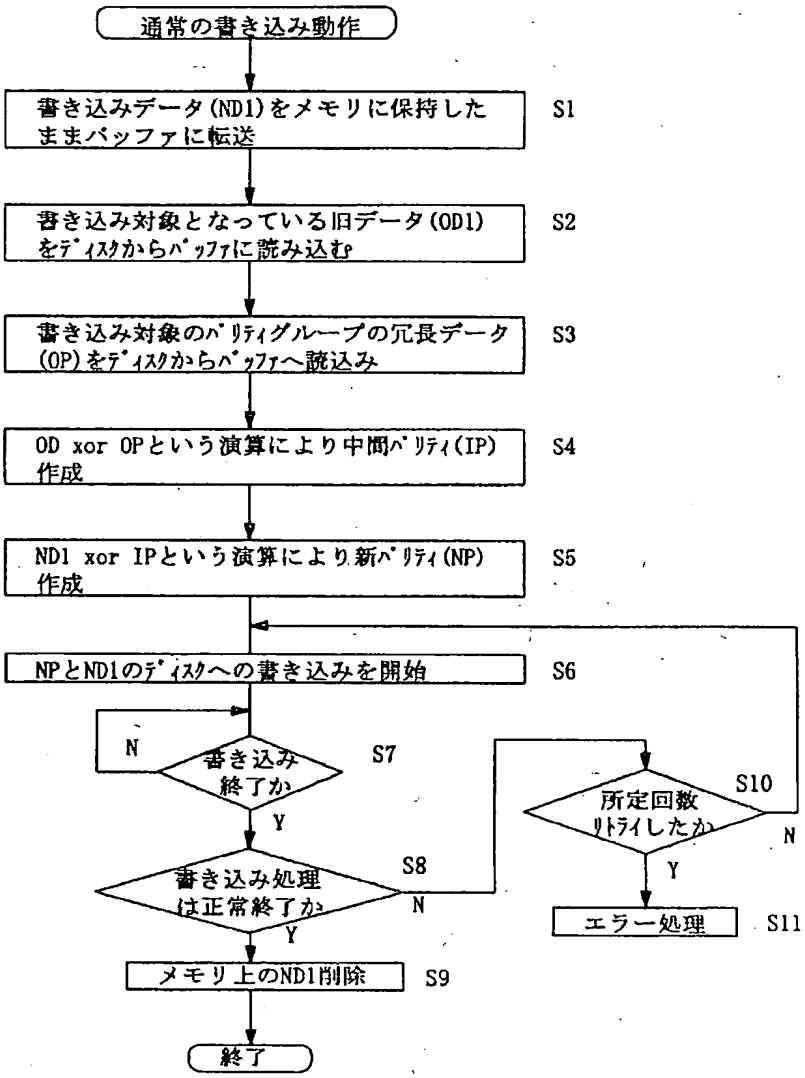
【図５】

縮退モードの設定処理のフローチャート



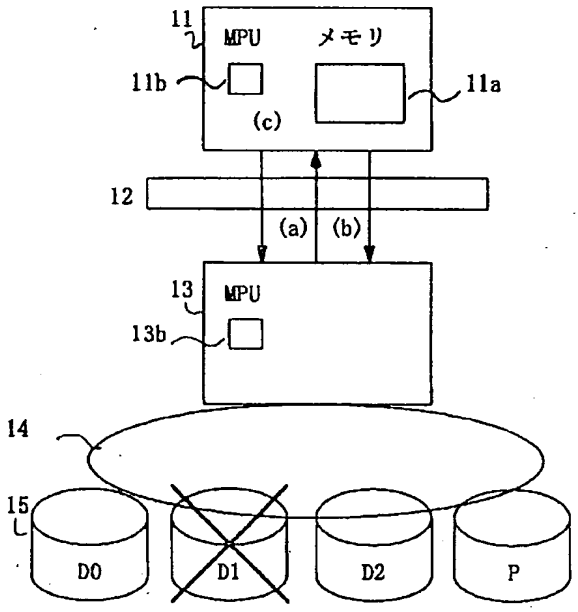
【図６】

正常時における書き込み処理のフローチャート



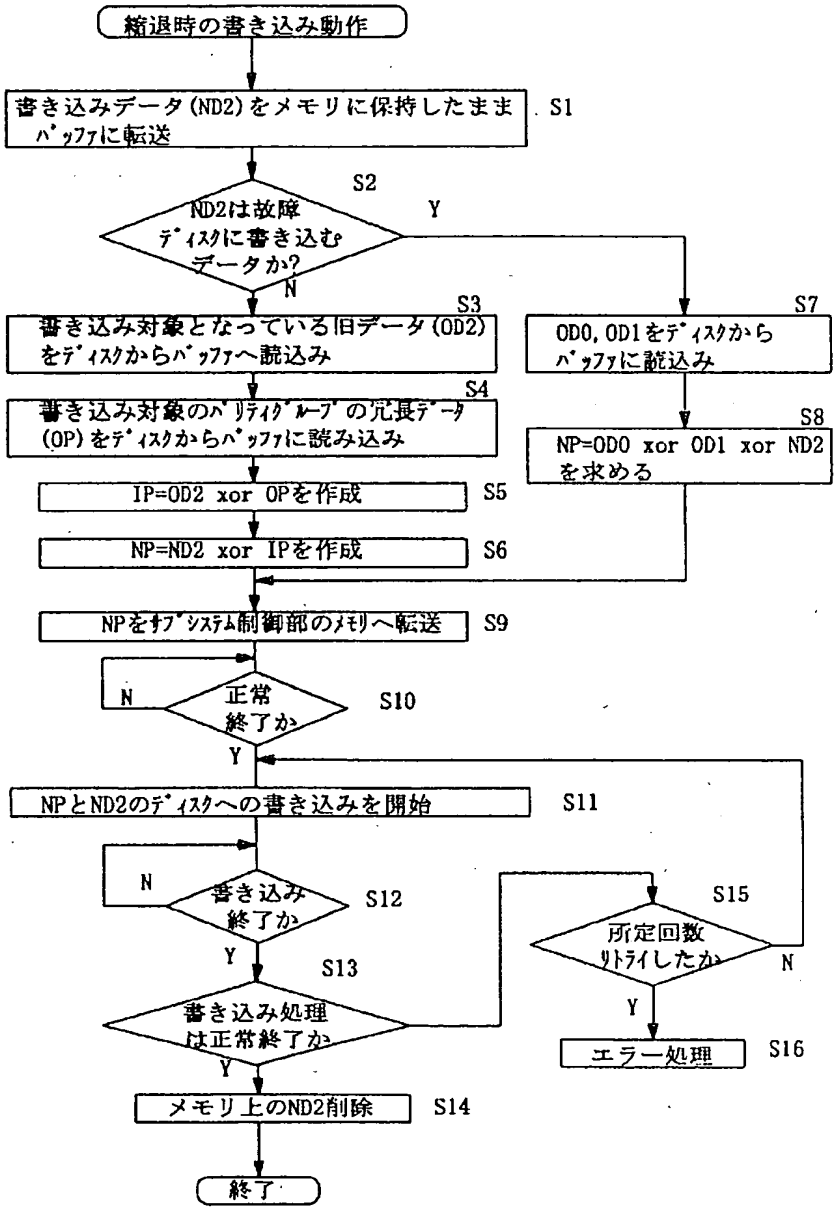
【図７】

ディスクドライブ群のあるディスクに何らかの異常が発生した場合の動作を説明する図



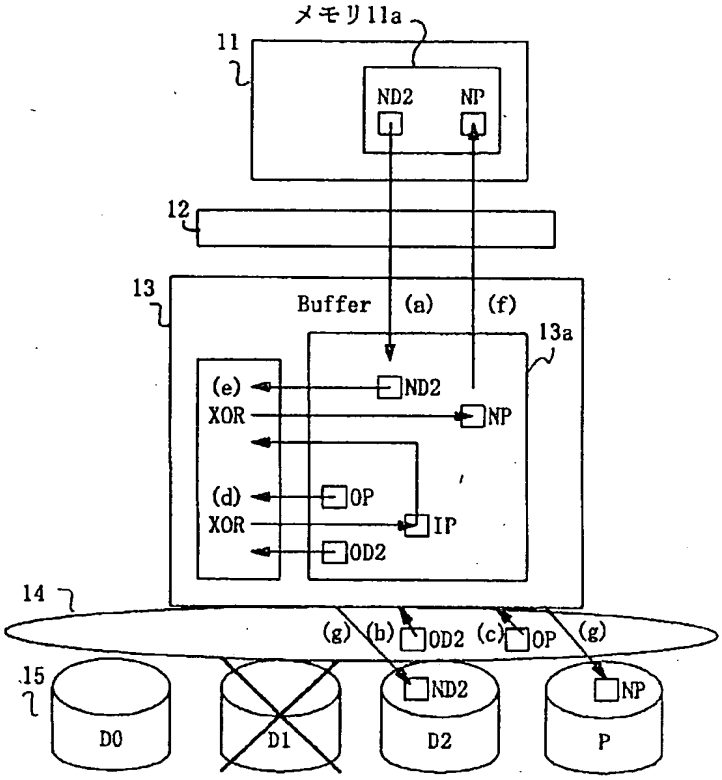
【図８】

縮退モードのときの書き込み動作を示すフローチャート



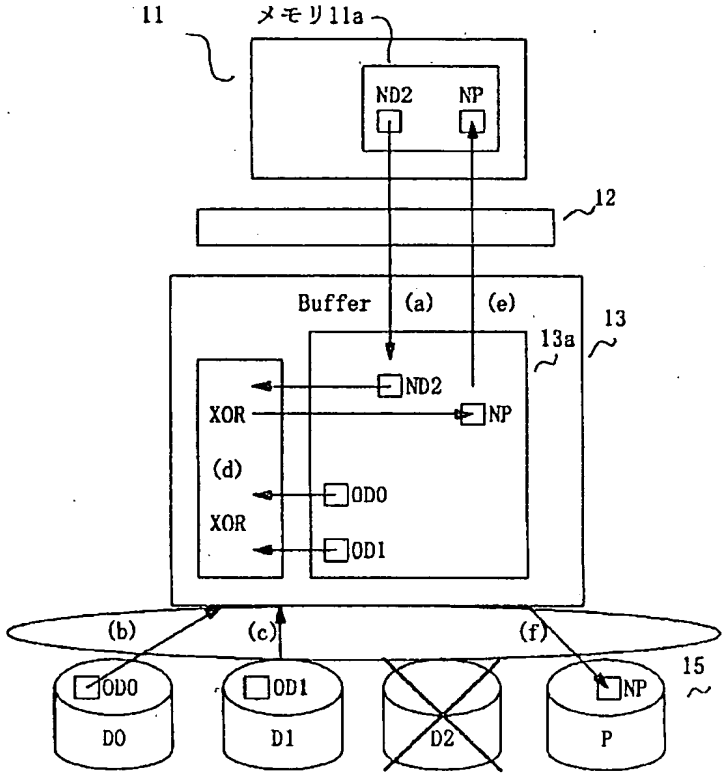
【図９】

縮退モードの書き込み動作シーケンス例（１）を示す図



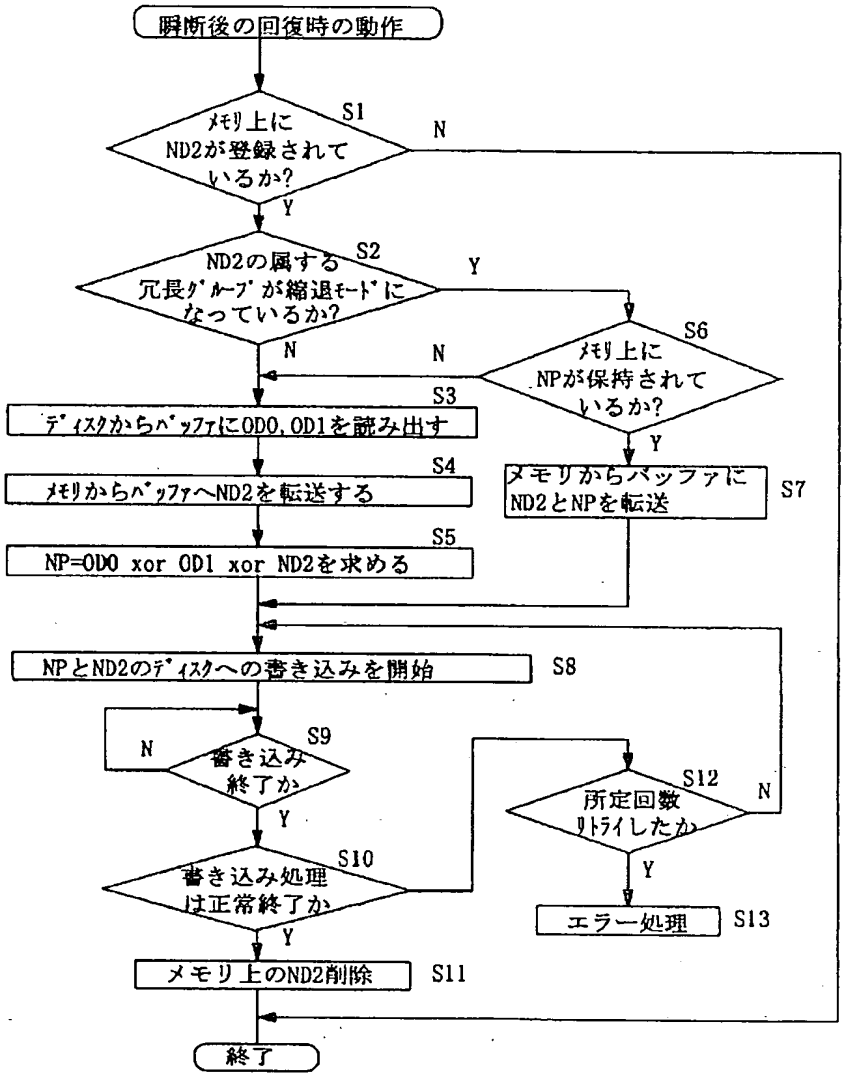
【図１０】

縮退モードの書き込み動作シーケンス例（２）を示す図



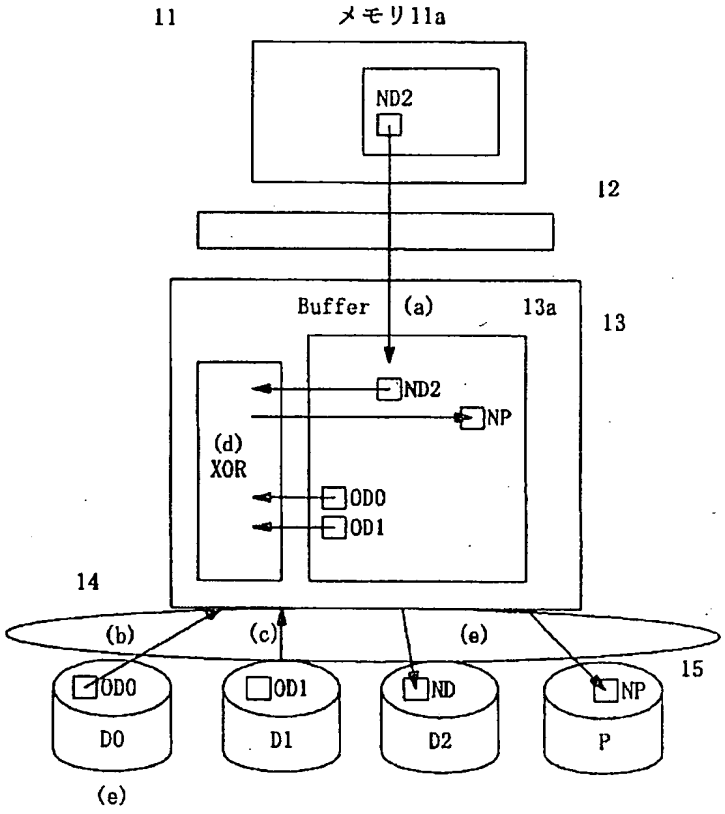
【図１１】

電源瞬断後の回復時の書き込み動作を示すフローチャート



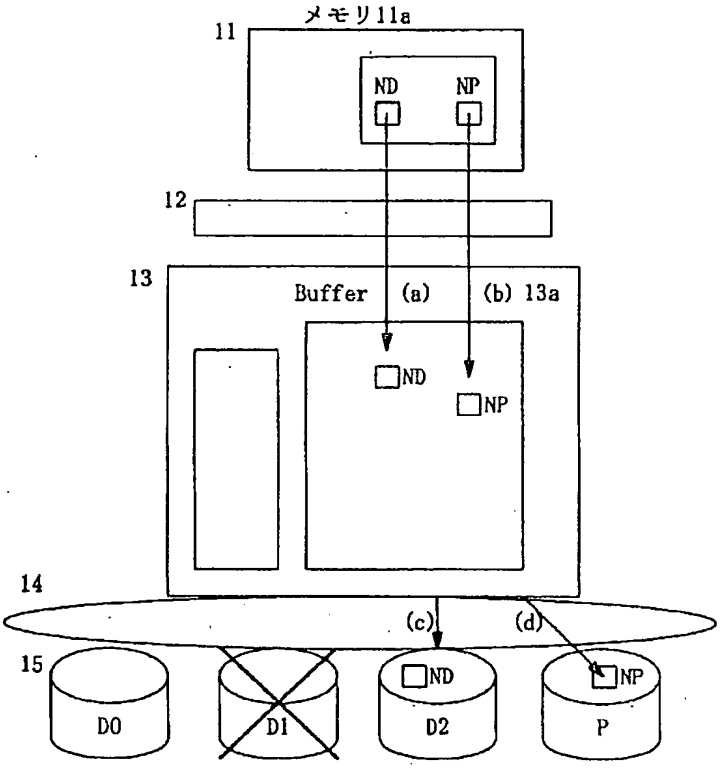
【図１２】

電源瞬断後の回復時の書き込み動作シーケンス例（１）を示す図



【図１３】

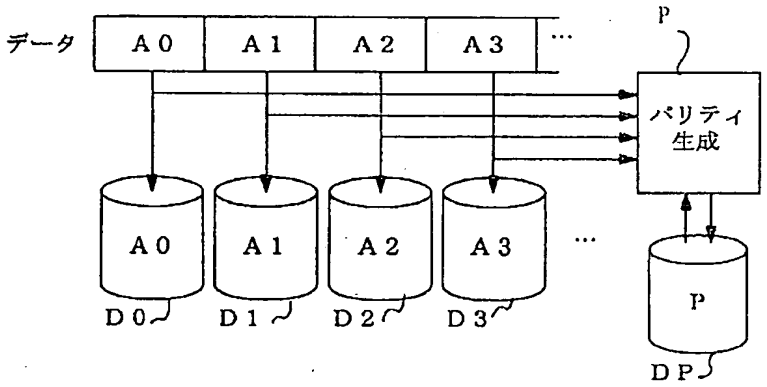
電源瞬断後の回復時の書き込み動作シーケンス例（２）を示す図



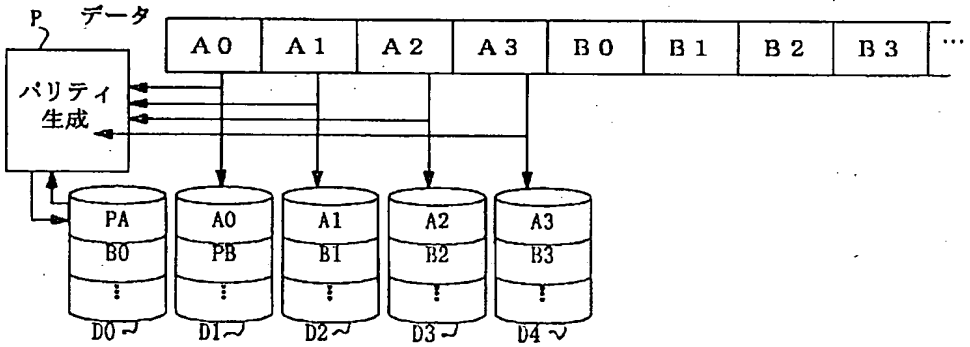
【図１４】

ＲＡＩＤ４，５を説明する図

(a)RAID4

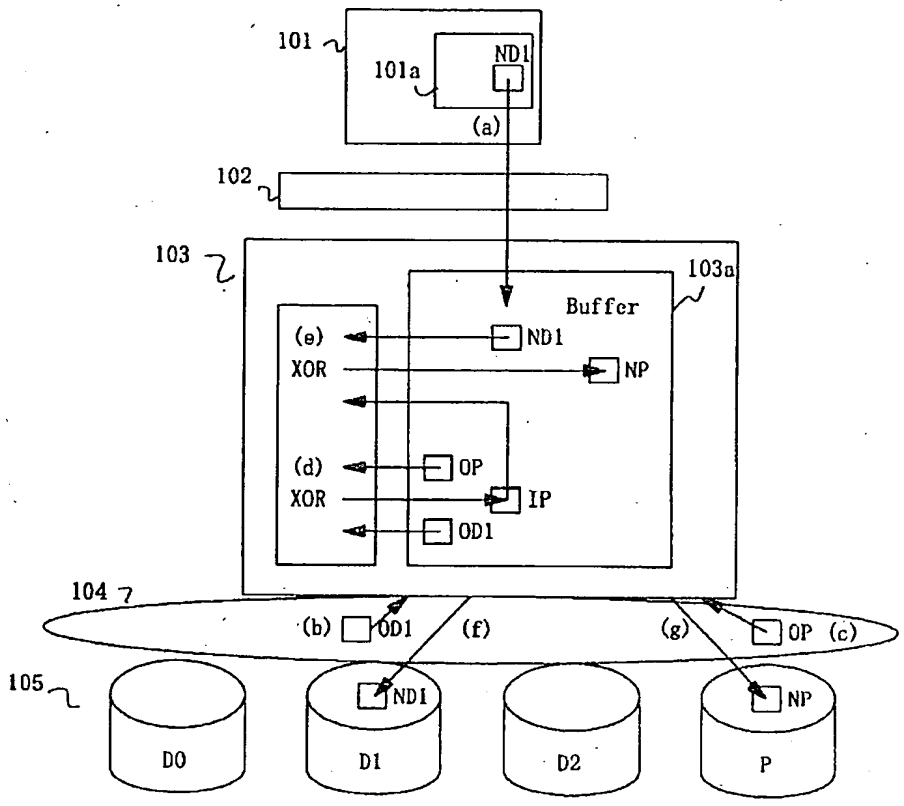


(b)RAID5



【図１５】

ディスクアレイ装置における書き込み動作シーケンスを示す図



フロントページの続き

(51)Int.Cl.[7]	識別記号	FI	テーマコード(参考)
G11B 20/18	570	G11B 20/18	570Z
	572		572B
			572F
	576		576Z

Fターム(参考)　5B018 GA02 KA15 KA21 LA06 MA14
QA05 QA06
5B065 BA01 CA11 CA30 CC08 CE12
EA02 EA23 EA24
5D044 BC01 CC04 DE68 DE94 HH07
HH17